

Instance-based and Feature-based Classification Enhancement for Short & Sparse Texts



Guodong Long

FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

UNIVERSITY OF TECHNOLOGY, SYDNEY

A thesis submitted for the degree of

Doctor of Philosophy

July, 2014

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student

Acknowledgements

First I would like to thank my supervisors, Prof Chengqi Zhang, Prof Xingquan Zhu and Dr Ling Chen. They brought me into the wonderful world of research. They not only gave me valuable advice in academics, but also helped my transition into a different culture. Prof Chengqi Zhang has been a great mentor and collaborator, energetic and full of ideas. Prof Xingquan Zhu guided me further into data mining. I am always impressed by his vigour and sharp thinking. Dr Ling Chen helped me by asking insightful questions, and giving me thoughtful comments on the thesis. I have enjoyed working with all of them, and have benefited enormously from the interactions with them.

I spent nearly four years at the University of Technology, Sydney. I thank the collaborators, faculty members, staff, fellow students and friends in the Centre for Quantum Computation and Intelligent Systems, who made my graduate life a very memorable experience. In particular, I thank you if you are reading this thesis.

Finally I thank my family. My parents endowed me with the gift of curiosity about the natural world. My sister, brother-in-law and parents-in-law gave me so much encouragement. Last but not least, my dear wife Jing brings to life so much love and happiness, making thesis writing an enjoyable endeavour.

Abstract

Short, sparse texts are becoming increasingly prevalent as a result of the growing popularity of social networking web sites, such as micro-blogs, Twitter and Flickr, and sites offering online product reviews. These short & sparse texts usually consist of a dozen or more words, or a few sentences, which we represent as a sparse document-term matrix. Compared to normal texts, short & sparse texts have three specific characteristics: (1) insufficient word co-occurrence to measure similarity, (2) low quality data resulting from spelling error, acronyms and slang, and (3) data sparseness. Normal classification methods therefore fail to achieve the desired level of accuracy for classifying short & sparse text.

In this thesis, we present a series of novel approaches to enhance the performance of short & sparse text classification. Most texts can be represented as a two-dimensional matrix and we use the terms - “instance” and “feature” to denote the “row” and “column” concept respectively in the matrix. Corresponding to the matrix’s two dimensions, we design an instance- and feature-based framework to expand the rows/columns in the matrix.

- for the instance-based framework, we extract an auxiliary dataset from an external online source (i.e. Wikipedia) with predefined class information, and integrate the target and auxiliary datasets with an instance-based transfer learning tool to enhance the classification performance of the target short text

domain. Moreover, we propose a sampling framework to handle the challenge of low quality data in auxiliary dataset;

- for the feature-based framework, we infer two kinds of feature sets with the given short texts, and then combine them with multi-view learning tool to enhance the classification performance. To handle the view disagreement challenge, we integrate a Bagging framework with Multi-view learning.

The aim of the proposed algorithms is to improve classification performance (i.e. accuracy). To evaluate the proposed algorithms, we test them using a variety of benchmark datasets and real world datasets, such as sentiment texts in Twitter, pre-processed 20 Newsgroup data, review texts for seminars, and search snippets. Moreover, we compare the algorithm with other benchmark algorithms on all datasets. The results of our experiments demonstrate that the accuracy of our proposed algorithms is superior to that of other similar algorithms.

Contents

Contents	v
List of Figures	ix
List of Tables	x
Nomenclature	x
1 Introduction	1
1.1 Background	1
1.2 Research Questions	3
1.3 Research Objectives	7
1.4 Significance and Main Contributions	10
1.5 Research Methodology	11
1.6 Thesis Structure	13
1.7 Publications Related to this Thesis	15
2 Literature Review	16
2.1 Applications of Short Texts	16
2.1.1 Query Snippet	17

2.1.2	Micro-blogs	17
2.1.3	Product Review, News, Blog Feeds, and Question Answering	18
2.1.4	Mobile SMS, Online chatting	18
2.1.5	TAG & Multi-media Description	18
2.2	Related Methods for Short Text Mining	19
2.2.1	Similarity Measurement of Short & Sparse Texts	19
2.2.2	Representative Methods in Short Texts Mining	21
2.2.2.1	Search Engine augmented Short Text Mining	21
2.2.2.2	Link Structure augmented Short Text Mining	22
2.2.2.3	Online Corpus augmented Short Text Mining	22
2.2.2.4	Words and Dictionary augmented Short Text Mining	23
2.2.2.5	Related Domain augmented Short Text Mining	23
2.2.2.6	Unlabelled Data augmented Short Text Mining	23
2.2.2.7	Other Sources augmented Short Text Mining	24
2.2.2.8	Dimension Reduction	24
2.2.2.9	Sparse Learner	24
2.2.2.10	Graph-based Semi-supervised Learning	24
2.2.2.11	Graph-based Methods	26
2.2.2.12	Complex Representation	27
2.2.2.13	Sentiment Analysis	27
2.3	Summary	27
3	Instance-based Classification Enhancement for Short & Sparse Texts	28
3.1	Introduction	29
3.2	The Proposed Method	31

3.2.1	Problem Definition	31
3.2.2	The Framework of TCSST	33
3.2.3	Data Preparation	34
3.2.4	Semi-supervised Classification	35
3.2.5	Transfer Classification	36
3.2.6	Sampling Unlabelled External Data	37
3.2.7	Boosting from Saved Classifiers	39
3.3	Summary	41
4	Feature-based Classification Enhancement for Short & Sparse Texts	42
4.1	Introduction	43
4.2	The Proposed Method	45
4.2.1	Definition	45
4.2.2	Framework of GMDS-MV	46
4.2.3	Construct Document Graph	47
4.2.4	Distance Matrix for Graph	47
4.2.5	Augment Feature Representation	48
4.2.6	Multi-View Classification	49
4.3	Summary	50
5	Performance Study and Experiment Results	52
5.1	Instance-based Classification Enhancement for Short & Sparse Texts	52
5.1.1	Datasets	53
5.1.2	Compared Methods	55
5.1.3	Experimental Results	57
5.1.3.1	Comparison between Algorithms.	57

5.1.3.2	Parameter Influence on TCSST	59
5.1.3.3	Iterations vs Accuracy of TCSST	60
5.1.3.4	Computational Complexity of TCSST	61
5.2	Feature-based Classification Enhancement for Short & Sparse Texts	61
5.2.1	Datasets	61
5.2.2	Visualizing the Graph-based MDS Representation	62
5.2.3	Compared Methods	63
5.2.4	Experimental Results	64
5.2.4.1	Accuracy Comparison	64
5.2.4.2	Parameter Influence on Algorithm	65
5.3	Summary	66
6	Conclusions and Future Research	67
6.1	Conclusions	67
6.2	Future Research	68
	Bibliography	70

List of Figures

1.1	Relationship between Chapters	14
3.1	The Framework of TCSST.	32
4.1	A Simple Example of Graph-based Short Text Classification.	44
4.2	Proposed Framework of GMDS-MV	46
5.1	Comparison of TCSST and Other Methods along a Wide Range of Data Variations	56
5.2	Parameter Influence on TCSST	59
5.3	Iterations vs Accuracy of TCSST w.r.t. MAX_ITERATION	60
5.4	PCA on Document-term Representation vs MDS on Graph Representation Sub- figures (a)(b)(c) Visualize 2-dimensional PCA based on Document-term Repre- sentaiont for 20newsgroup/Twitter Sentiment/Search Snippet Dataset Subfig- ures (d)(e)(f) Visualize 2-dimensional MDS based on Document Graph Repre- sentation for 20newsgroup/Twitter Sentiment/Search Snippet Dataset	62
5.5	Accuracy Changes on Various Parameters in Search Snippet Data	65

List of Tables

5.1	dataset constructed from 20-Newsgroup	54
5.2	Statistics of the two data sets	54
5.3	Accuracy of TCSST and other methods on the benchmark data	55
5.4	Accuracy of TCSST and other methods on the NRM data	58
5.5	Accuracy comparison on three datasets	65